## ABSTRACT

The present invention relates to a methodology for assembling a document from content spanning multiple web-pages employing two cooperative processes. Given a starting location, one process analyzes a single page at a time to find candidate links. The links are recursively followed and those pages are analyzed. A detailed set of heuristics is used to determine what is or is not a candidate link. The links are examined for link clusters and a table of contents if found is identified. The candidate pages are then fed to a document-level analyzer. This process compares the attributes of one page against the others and looks for a document-like structure. Using another detailed set of heuristics, the document-level analyzer determines if the page should be included in the document.